Some reflections on data in the public sector...

Tom Moritz

Internet Archive



London School of Economics
London
March 26-27, 2009

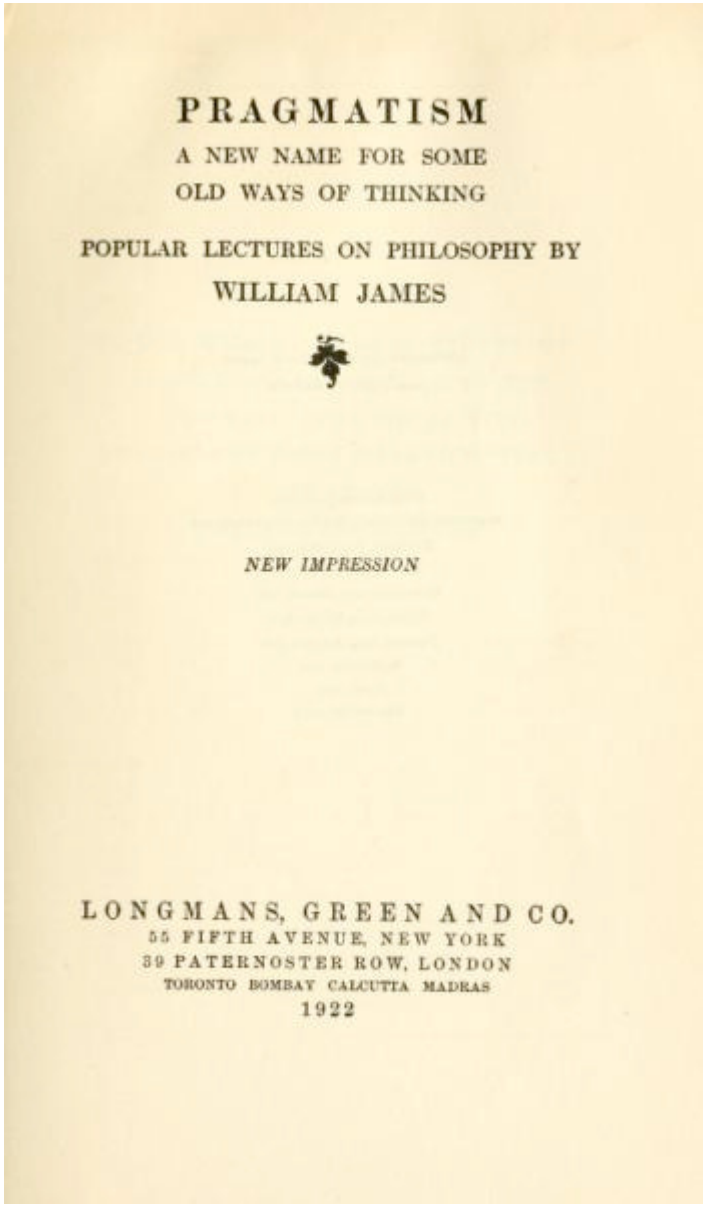
COMMUNIA
The European Thematic Network on the Digital Public Domain

“Data”?

Clear definitions are good. We should not rely on metaphysical “solving” / “power-bringing” words:

“So the universe has always appeared to the natural mind as a kind of enigma, of which the key must be sought in the shape of some illuminating or power-bringing word or name. That word names the universe's principle, and to possess it is, after a fashion, to possess the universe itself
'God,' 'Matter,' 'Reason,’
'the Absolute,’ ‘Energy,’
[“Knowledge” / “Information” / “Data” -- added]
are so many solving names.
You can rest when you have them. You are at the end of your metaphysical quest.”

William James. "What Pragmatism Means". Lecture 2 in Pragmatism: A new name for some old ways of thinking. New York: Longman Green and Co (1922): 52-52.
http://www.archive.org/stream/pragmatismnewnam00jame



PRAGMATISM

A NEW NAME FOR SOME
OLD WAYS OF THINKING

POPULAR LECTURES ON PHILOSOPHY BY
WILLIAM JAMES

NEW IMPRESSION

LONGMANS, GREEN AND CO.
55 FIFTH AVENUE, NEW YORK
39 PATERNOSTER ROW, LONDON
TORONTO BOMBAY CALCUTTA MADRAS
1922

Internet Archive:
http://www.archive.org/stream/pragmatismnewnam00jame

Note Date of Publiction: ***1922***

Data?

In common usage “data” refers both

to an electronic medium of exchange (this is the definition applied by the US NSF “DataNet” solicitation)

And

disciplinarily / epistemically it refers to formal, consistent, conventional expressions of facts (observations/ measurements)

We should be clear how we are using the term.

[BTW in normal usage “data” can be singular or plural…?]

Digital Explosion

- "The digital universe in 2007 — at 2.25 x 1021bits (281 exabytes or 281 billion gigabytes) — was 10% bigger than we thought. The resizing comes as a result of faster growth in cameras, digital TV shipments, and better understanding of information replication.

- "By 2011, the digital universe will be 10 times the size it was in 2006.

- "As forecast***, the amount of information created, captured, or replicated exceeded available storage for the first time in 2007***. Not all information created and transmitted gets stored, but by 2011, almost half of the digital universe will not have a permanent home."

The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth through 2011. An IDC Whitepaper
www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Data is now more than ever available
in highly diverse formats from very disparate sources:


SAFETY BUTTON POPS UP WHEN ORIGINAL SEAL IS BROKEN
SAFETY BUTTON
Snapple
POPS UP
REJECT IF BUTTON IS UP
"Real Fact" #134
An ant can lift 50 times its own weight.
Get all the "Real Facts" at snapple.com

Validation of data and critical awareness and analysis of data sources is essential.

- We must address both *legacy data* and *current/ prospective data*
- Many data sets – to be fully useful must be significantly *longitudinal* – for example – biological taxonomy – but also climate, oceanography, etc
- Older data sets while essential may be much more problematic
  - Russian Chronicles of Nature / zapovedniks
  - US LTER Trout Lake, WI example – (Geof Bowker)
  - California Fish & Game

NCAR Research Data Archive (RDA)

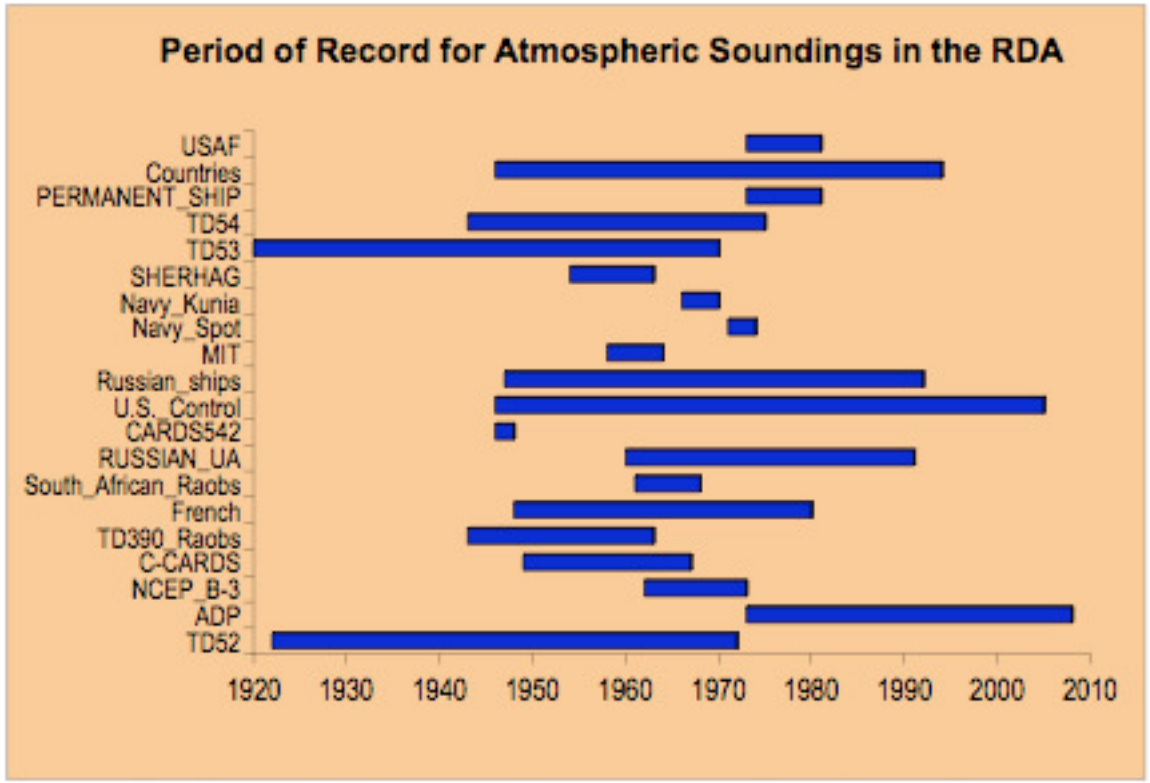


Figure 2 Period of record for major sources of atmospheric sounding data in the NCAR Research Data Archive. The name of each dataset source is shown along with date range covered.

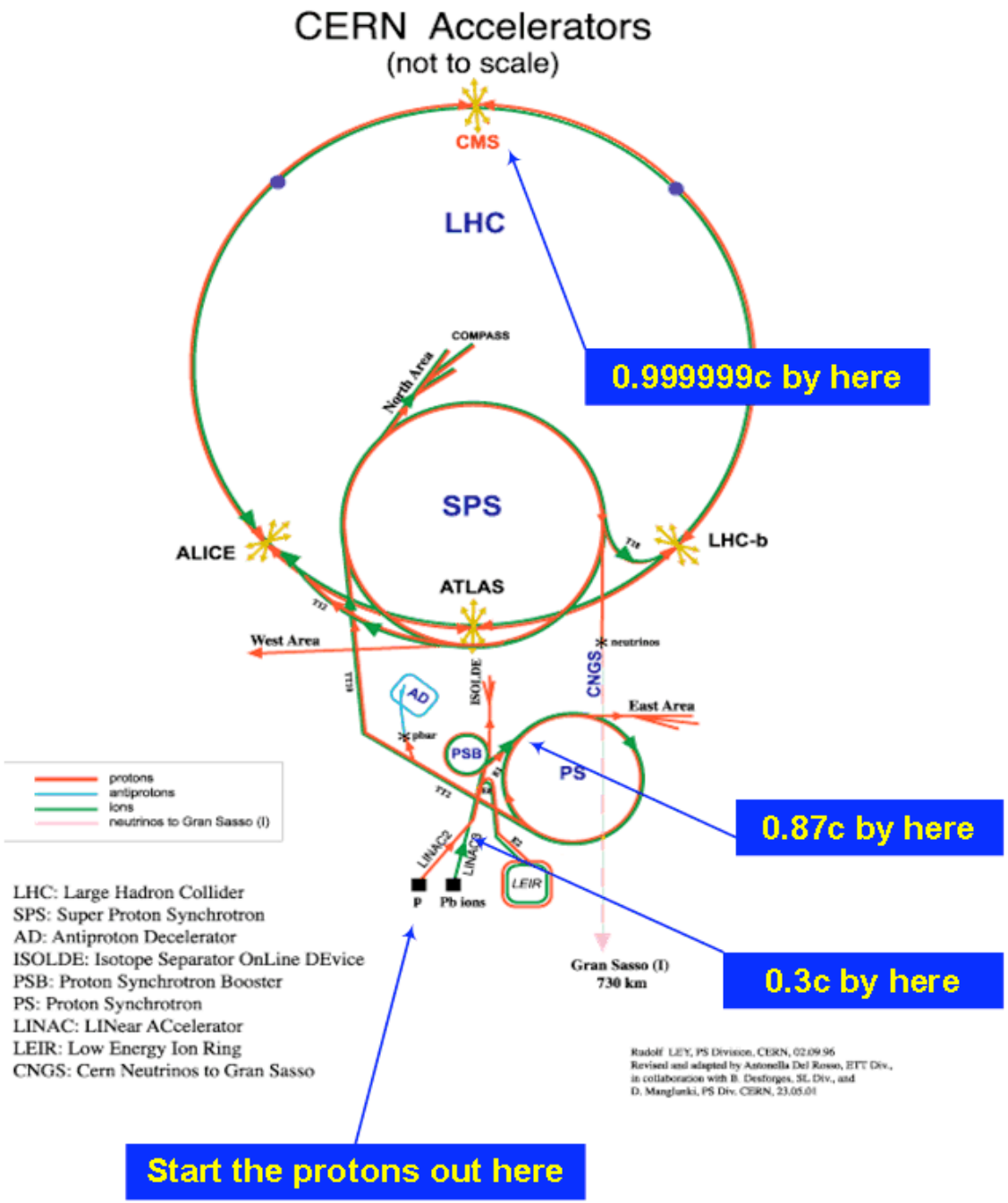
C.A. Jacobs, S. J. Worley, *"Data Curation in Climate and Weather: Transforming our ability to improve predictions through global knowledge sharing ,"* from the **4th International Digital Curation Conference December 2008 , page 7.** www.dcc.ac.uk/events/dcc-2008/programme/papers/Data%20Curation%20in%20Climate%20and%20Weather.pdf [03 02 09]

The NCAR Research Data Archive (RDA)

"The NCAR Research Data Archive (RDA) is a comparatively small (currently 246 TB, less than 5% of the MSS [Mass Storage System] total size), but very important, part of the MSS stored data. The RDA has been *curated by the staff in the Computational and Information Systems Laboratory for over 40 years*, [emphasis added] and as such contains reference datasets used by large numbers of scientists. The RDA contents are long-term atmospheric (surface and upper air) and oceanographic observations, grid analyses of observational datasets, operational weather prediction model output, reanalyses, satellite derived datasets, and ancillary datasets, such as topography/bathymetry, vegetation, and land use. The RDA is not a static collection; *it is now over 580 datasets* with about *100 routinely updated* and *10-20 new ones added each year*. "

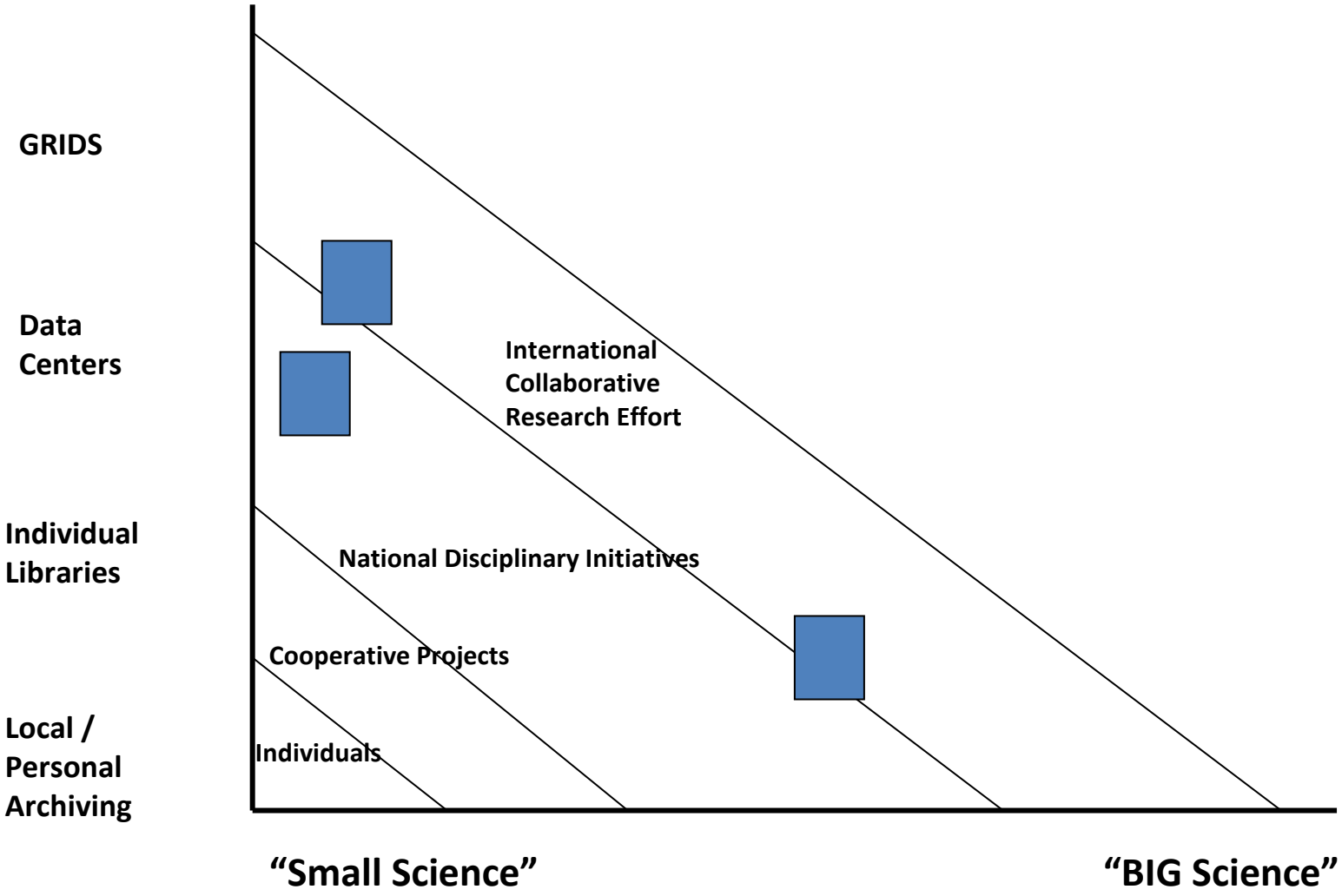
C.A. Jacobs, S. J. Worley, *"Data Curation in Climate and Weather: Transforming our ability to improve predictions through global knowledge sharing ,"* from the **4th International Digital Curation Conference December 2008, page 5.** www.dcc.ac.uk/events/dcc-2008/programme/papers/Data%20Curation%20in%20Climate%20and%20Weather.pdf [03 02 09]

- In some instances we are working at “peta-scale”
- This has dramatic implications for future full-life cycle management
- quantity becomes quality?



The $3.6 billion Large Hadron Collider (LHC) will sample and record the results of up to 600 million proton collisions per second, producing roughly 15 petabytes (15 million gigabytes) of data annually in search of new fundamental particles. To allow thousands of scientists from around the globe to collaborate on the analysis of these data over the next 15 years (the estimated lifetime of the LHC), tens of thousands of computers located around the world are being harnessed in a distributed computing network called the Grid. Within the Grid, described as the most powerful supercomputer system in the world, the avalanche of data will be analyzed, shared, re-purposed and combined in innovative new ways designed to reveal the secrets of the fundamental properties of matter.

LHC source:
http://public.web.cern.ch/public/en/LHC/LHC-en.html
Source:
http://public.web.cern.ch/Public/en/LHC/LHC-en.html


GRIDS
Data Centers
Individual Libraries
Local / Personal Archiving
International Collaborative Research Effort
National Disciplinary Initiatives
Cooperative Projects
Individuals
"Small Science"
"BIG Science"

“Small Science” DATA SETS

some examples with “native metadata”

2-d_soil_temps.csv

surface, and sub-surface soil temperatures (at 2cm and 8cm depths) measured at one location for a few days in order to calibrate a model of temperature propagation. Surface temperature was measured with an infrared thermometer, subsurface temperatures with a thermocouple.

5-minute_light_data_for_4_continuous_days_plus_reference.xls

PPF (photosynthetic photon flux = photosynthetically active radiation 400-700nm) measured with an array of photodiodes calibrated to a Licor sensor, along a linear transect for a few days. used to get an idea of how much light plants along the transect are receiving.

CO2_of_air_at_different_heights_July_9.xls

concentration of CO2 in the air during the evening for one day, measured with a Licor infrared gas analyzer and a series of relays and tubes with a pump. used to examine the gradient of CO2 coming from the soil when the air is still during the evening.

Fern_light_response.xls

Light response curves for bracken ferns, measured with a Licor photosynthesis system. Fronds are exposed to different light levels and their instantaneous photosynthesis and conductance is measured. used in conjunction with the induction data (below) for physiological characterization of the ferns.

La_Selva_species_photosyntheis_table.xls

incomplete data set on instantaneous photosynthesis rates for various tropical understory and epiphytic species grown in a shade house in Costa Rica.

---

manzanita_sapflow_12-5-07_to_7-7-08.xls

instantaneous sap flow data (as temperature differences on a constant temperature heat dissipation probe) for multiple branches of Manzanita, collected with a datalogger. used to correlate physiological activity with below-ground measures of root grown and CO2 production.

---

moisture_release_curves.xls

percentage of water content, water potential (in MegaPascals) and temperature of soil samples, measured in the laboratory for calibration of water content with water potential. soil is from the James Reserve in California.

Photosynthetic_induction.xls

a time-course of photosynthetic induction for a leaf over 35 minutes. instantaneous photosynthesis measured as ❒mol CO2 m/2/s and light level is probably 1000 micromoles. used to determine physiological characteristics of bracken ferns.

---

run_2_24-h_data_for_mesh.xls

measurements of micrometeorological parameters on a moving shuttle, going from a clearing across a forest edge and into the forest for about 30 meters. Pyronometers facing up and down, pyrgeometer facing up and down, PAR, air temperature, relative humidity. Also data from a station fixed in the clearing and some derived variables calculated. used for examining edge effects in forests.

Segment_of_wallflower_compare_colorspaces_blur.xls

pixel counts from images of wallflowers that were segmented into flower/not-flower under different color spaces. segmentation was made using a probability matrix of hand-segmented images. used to automatically count flowers in images collected after this training data was collected (and used to determine the best color space for this task).

manzanita_sapflow_12-5-07_to_7-7-08.xls

instantaneous sap flow data (as temperature differences on a constant temperature heat dissipation probe) for multiple branches of Manzanita, collected with a datalogger.
used to correlate physiological activity with below-ground measures of root grown and CO2 production.

sbid	*battery*	*datetime*	*heater_voltage*	*Manz1Sap1*	*Manz1Sap2*	*Manz1Sap3*	*Manz1Sap4*	*Manz2Sap5*	*Manz2Sap6*	*Manz2Sap7*	*Manz3Sap10*	*Manz3Sap8*	*Manz3Sap9*	*Manz4Sap11*	*timestamp*	*Datagap*	*Julian*
2	12.365	1196796112		2018.8	0.5585	0.51029	0.55517	0.54354	0.6067	0.52858	0.55351	0.59008	0.59506	0.60337	0.56514	12/4/07 11:21		4.47351
3	12.348	1196796232		2017.9	0.55682	0.51028	0.5535	0.54352	0.60669	0.52857	0.55017	0.59007	0.59505	0.60336	0.56513	12/4/07 11:23	0	4.47490
4	12.357	1196796352		2018.6	0.55514	0.51027	0.55348	0.54351	0.60501	0.52855	0.55016	0.59005	0.59504	0.60501	0.56512	12/4/07 11:25	0	4.47628
5	12.354	1196796472		2017.6	0.55514	0.51026	0.55181	0.5435	0.60334	0.52855	0.54849	0.59004	0.59503	0.60334	0.56511	12/4/07 11:27	0	4.47767
6	12.334	1196796592		2018.3	0.55347	0.51026	0.55015	0.5435	0.60333	0.52854	0.54682	0.59004	0.59502	0.605	0.56511	12/4/07 11:29	0	4.47906
7	12.34	1196796712		2018.5	0.55014	0.50859	0.55014	0.54349	0.60332	0.53019	0.54349	0.59003	0.59501	0.60498	0.56676	12/4/07 11:31	0	4.48045
8	12.337	1196796832		2017.8	0.55013	0.50692	0.55013	0.54348	0.60332	0.53019	0.54182	0.59002	0.59501	0.60498	0.56675	12/4/07 11:33	0	4.48184
9	12.328	1196796952		2017.5	0.5468	0.50691	0.5468	0.54347	0.60331	0.53018	0.53849	0.59001	0.595	0.60497	0.56674	12/4/07 11:35	0	4.48323
10	12.323	1196797072		2017	0.54679	0.50524	0.54679	0.54347	0.59998	0.53017	0.53682	0.59	0.59499	0.60496	0.56674	12/4/07 11:37	0	4.48462
11	12.328	1196797192		2018.9	0.54679	0.50191	0.54512	0.5418	0.5966[illegible]	0.53017	0.53349	0.59	0.59498	0.60496	0.56673	12/4/07 11:39	0	4.48601
12	12.319	1196797312		2017.7	0.54345	0.49857	0.54178	0.54178	0.59663	0.53015	0.53015	0.58998	0.5933	0.60327	0.56671	12/4/07 11:41	0	4.48740
13	12.311	1196797432		2017.3	0.54343	0.4969	0.54011	0.54177	0.59661	0.53014	0.52848	0.58997	0.59329	0.6016	0.5667	12/4/07 11:43	0	4.48878
14	12.316	1196797552		2018.6	0.5401	0.49357	0.53678	0.54176	0.59328	0.53013	0.5268	0.58995	0.59328	0.60325	0.56669	12/4/07 11:45	0	4.49017
15	12.31	1196797672		2016.8	0.53844	0.4919	0.53511	0.54176	0.59494	0.53013	0.52514	0.58995	0.59328	0.60325	0.56503	12/4/07 11:47	0	4.49156
16	12.31	1196797792		2017.1	0.53676	0.48856	0.53343	0.54174	0.59326	0.53011	0.5218	0.58993	0.59326	0.60323	0.56501	12/4/07 11:49	0	4.49295
17	12.31	1196797912		2017.1	0.53342	0.48523	0.5301	0.54173	0.59324	0.5301	0.51846	0.58826	0.59324	0.60321	0.56499	12/4/07 11:51	0	4.49434
18	12.301	1196798031		2017.5	0.53174	0.48521	0.52842	0.53839	0.59156	0.53008	0.51845	0.58824	0.59323	0.6032	0.56498	12/4/07 11:53	0	4.49573
19	12.301	1196798151		2016.3	0.53007	0.48188	0.52509	0.53838	0.59155	0.53007	0.51512	0.58823	0.59321	0.60152	0.5633	12/4/07 11:55	0	4.49712
20	12.303	1196798271		2016.6	0.5284	0.47855	0.52175	0.53837	0.59154	0.5284	0.51[illegible]1	0.58821	0.59154	0.60151	0.56163	12/4/07 11:57	0	4.49851

A Datum: “0.59998”

From an Excel Spreadsheet

- *Access per se* does not equal *fitness for use*
- *Provenance or context* are essential to give meaning to data [SEE February letter to *Science: "Keeping Raw Data in Context"*]
- Geo-scale is of particular importance for PSI
- Mechanisms for maintaining provenance -- through combinations and re-combinations of data -- are essential
- GBIF in Copenhagen has recently formed a *Data Publishing Framework Task Group*

GBIF – October, 2008

(as a result of the Darwin Core reductionist data analysis…)

GBIF UDDI Registry

* registration
* update information

Data Providers **259**

Datasets **7481**

Searchable Records **147,539,975**

http://www.gbif.org/ [clipped Oct 8, 2008]

- Data does not respect sectors – it is easy to envision integral data sets drawn from public / private for-profit / not-for-profit sectors
- At a recent US NAS hearing the Dow Chemical Company reported that it had several hundred thousand technical reports in a proprietary corporate collection…
  - The greatly extended latency (?) of public access to this work is a violation of a fundamental principle of science
- We must exert pressure for free/ open access and use *in all domains (Wellcome Trust has been exemplary)*

Sakhalin Energy relocates offshore pipelines to protect whales



30/03/2005

"Yuzhno-Sakhalinsk, Russian Federation, 30 March 2005: Sakhalin Energy will reroute offshore pipelines in its oil and gas development in the Russian Far East to help protect the endangered western gray whale. "

http://www.shell.com/home/content/media/news_and_library/press_releases/2005/sakhalin_energy_relocates_pipeline_30032005.html

We do not know how data might be used and who might use it…



Evolution and Ecology of the Digital Domain

Stages of Digital Library Development

Stage	Date	Sponsor	Purpose
I: Experimental	1994	NSF/ARPA/NASA	Experiments on collections of digital materials
II: Developing	1998/1999	NSF/ARPA/NASA, DLF/CLIR	Begin to consider custodianship, sustainability, user communities
III: Mature	?	Funded through normal channels?	Real sustainable interoperable digital libraries

Howard Besser. Adapted from *The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries* by First Monday, volume 7, number 6 (June 2002),
URL: http://firstmonday.org/issues/issue7_6/besser/index.html

"the institutional ecology of the digital environment" (Yokai Benkler)

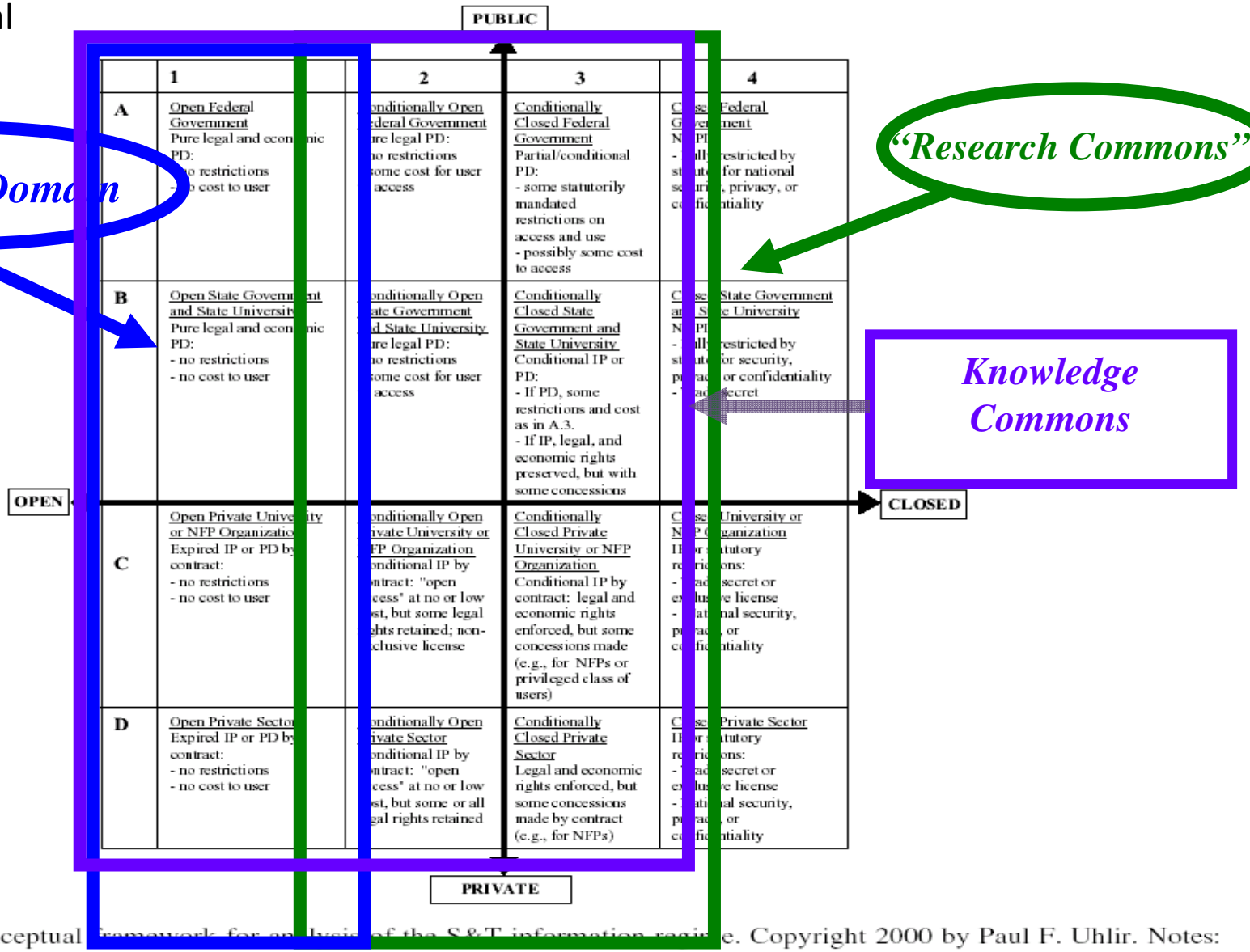
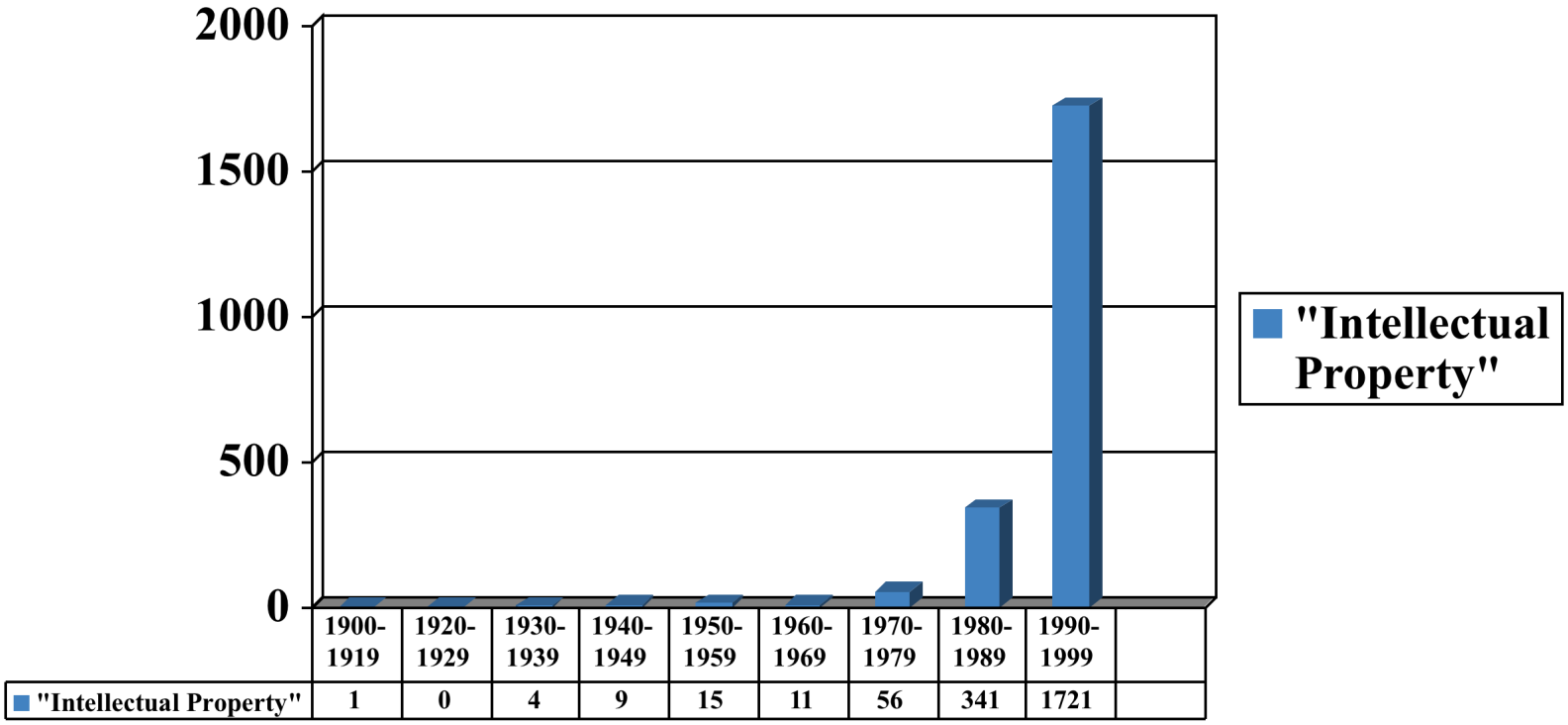


FIGURE 1-1 Conceptual framework for analysis of the S&T information regime. Copyright 2000 by Paul F. Uhlir. Notes: "PD"-public domain; "IP"-intellectual property; "NFP"-not-for-profit.

THE ROLE OF SCIENTIFIC AND TECHNICAL DATA AND INFORMATION IN THE PUBLIC DOMAIN PROCEEDINGS OF A SYMPOSIUM **Julie M. Esanu and Paul F. Uhlir, Editors Steering Committee on the Role of Scientific and Technical Data and Information in the Public Domain Office of International Scientific and Technical Information Programs Board on International Scientific Organizations Policy and Global Affairs Division, National Research Council of the National Academies, p. 5**

References to "*Intellectual Property*" in U.S. federal cases



	1900-1919	1920-1929	1930-1939	1940-1949	1950-1959	1960-1969	1970-1979	1980-1989	1990-1999	
"Intellectual Property"	1	0	4	9	15	11	56	341	1721	

"Professor Hank Greely" Cited in Lessig, L. *The future of ideas: the fate of the commons in a connrcted world.* NY, Random House, 2001. P. 294.

Graph of "The Knowledge Life Cycle"

Julian Birkinshaw and Tony Sheehan, "Managing the Knowledge Life Cycle,"
MIT Sloan Management Review, 44 (2) Fall, 2002: 77.

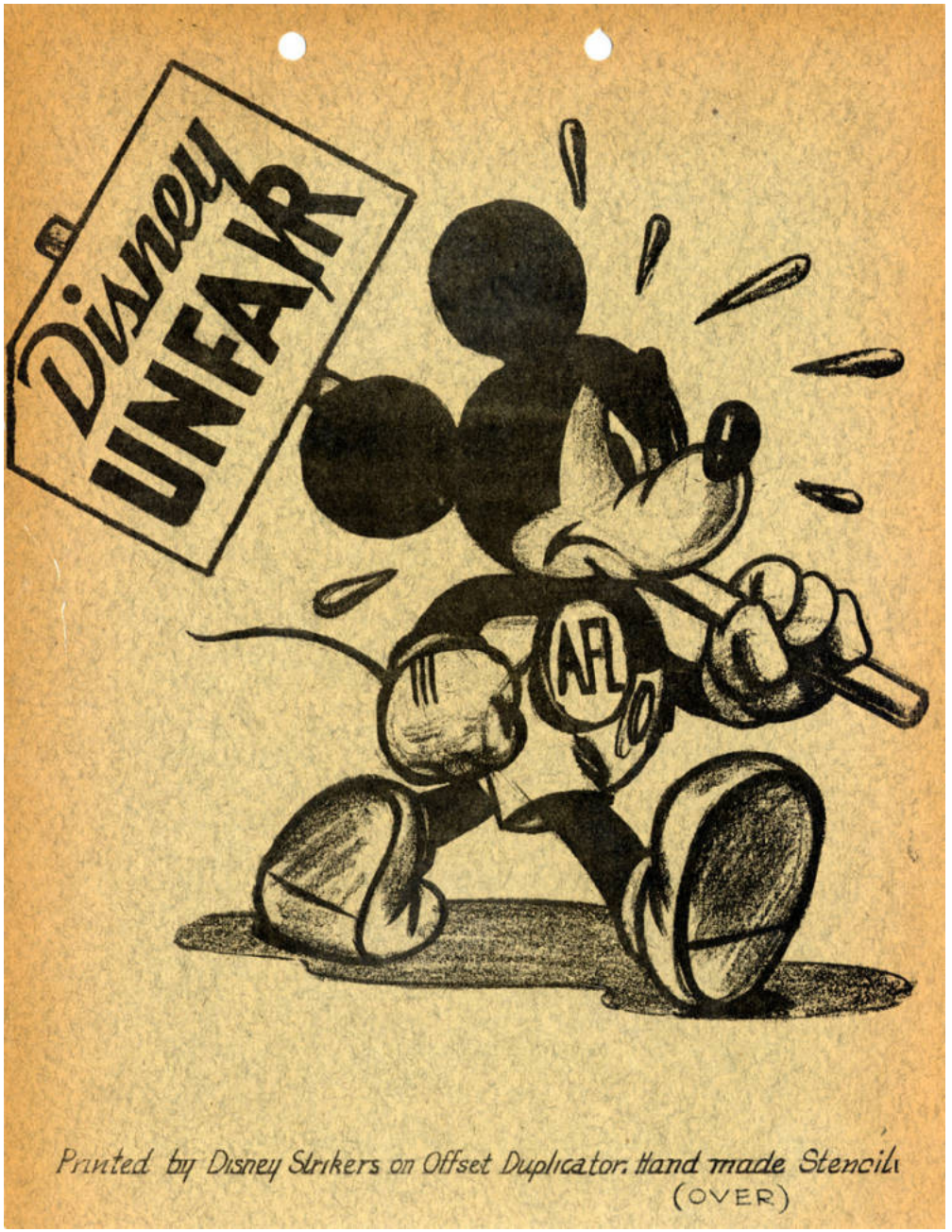
Shows: "Creation → Mobilization → Diffusion → Commoditization" of knowledge as developmental cycle over time with "access" increasing significantly to final "commoditization" stage….

Added annotation: "Is scientific knowledge a commodity ?"

[Flier from 1941 cartoonists strike at Disney Studios]

“Mickey Mouse wears an AFL (American Federation of Labor) button and carries a placards that reads "Disney UNFAIR." Bottom edge reads ‘Printed by Disney Strikers on Offset Duplicator. Hand made Stencil’ “

[*Metadata:* Strikes and lockouts -- Motion picture industry; Walt Disney Productions; Disney characters; Mickey Mouse; Motion picture industry -- Employees -- Labor unions; American Federation of Labor; Animators; Brotherhood of Painters, Decorators, and Paperhangers of America.; Screen Cartoonists Local Union No. 852 (Hollywood, Calif.); Animation Guild and Affiliated Optical Electronic and Graphic Arts, Local 839 I.A.T.S.E. (North Hollywood, Los Angeles, Calif.); Motion Pictures Screen Cartoonists Local 839, I.A.T.S.E.]



Cal State Univ Northridge

http://digitallibrary.csun.edu/cdm4/results.php?CISOOP1=any&CISOBOX1=Disney&CISOFIELD1=CISOSEARCHALL&CISOROOT=all&submit=search

Perhaps certain types
of “cultural properties”
are *inevitably*
commodities?

- Perhaps some cultural objects and works WITH HIGH MARKET VALUE will inevitably fall into restricted use [art , talkies, vampire novels…?] – but much work – including orphaned works and out-of-print work [demonstrably non-commercial] should be available for access and use
- The groups that sued Google are representative of major commercial interests
- The “long tail” case seems convincing but we must consider the societal cost-benefit analysis that leads from it to severe restrictions on access in exchange for very marginal cost-benefits to individual producers
- Perhaps some simple one-time opt-out, opt-in or buy-out?
- Or as Jonas Salk noted the reward is the ability to go on and to do more…

- Libraries, archives and museums have – for better or worse – long been the accepted repositories for human knowledge…
- The notion of commercial "corporations" serving as custodians of knowledge is highly problematic
- A problem of mission –
  - MicroSoft made a business decision last year (2008) to stop digital activity
  - the oldest known human corporation [Japan's Kongō Gumi -- a construction company founded in 578] was sold and consolidated into another company

- Monopolies and cartels are bad – Elsevier ?
  - In 2004 in the Washington Post Elsevier reported a 34% profit margin
- But they are clever (“smartest guys in the room”? -- ENRON? AIG? ….Google?)

There are powerful, well-formed arguments for the contributions open access and effective use of data to the public welfare.

These arguments are drawn from notions of:

Human rights / Fairness
Secular democracy
Civic responsibility
The ethos of science
The ethos of conservation
Education / Scientific literacy
Public health
And others…

“If Avian Flu Has Passed Us By Here’s Why…” (NYT)

- Chart showing global spread of avian flu – together with hemispheric avian migration routes…

- Text added:
- “How many data sources contributed to this analysis…?”

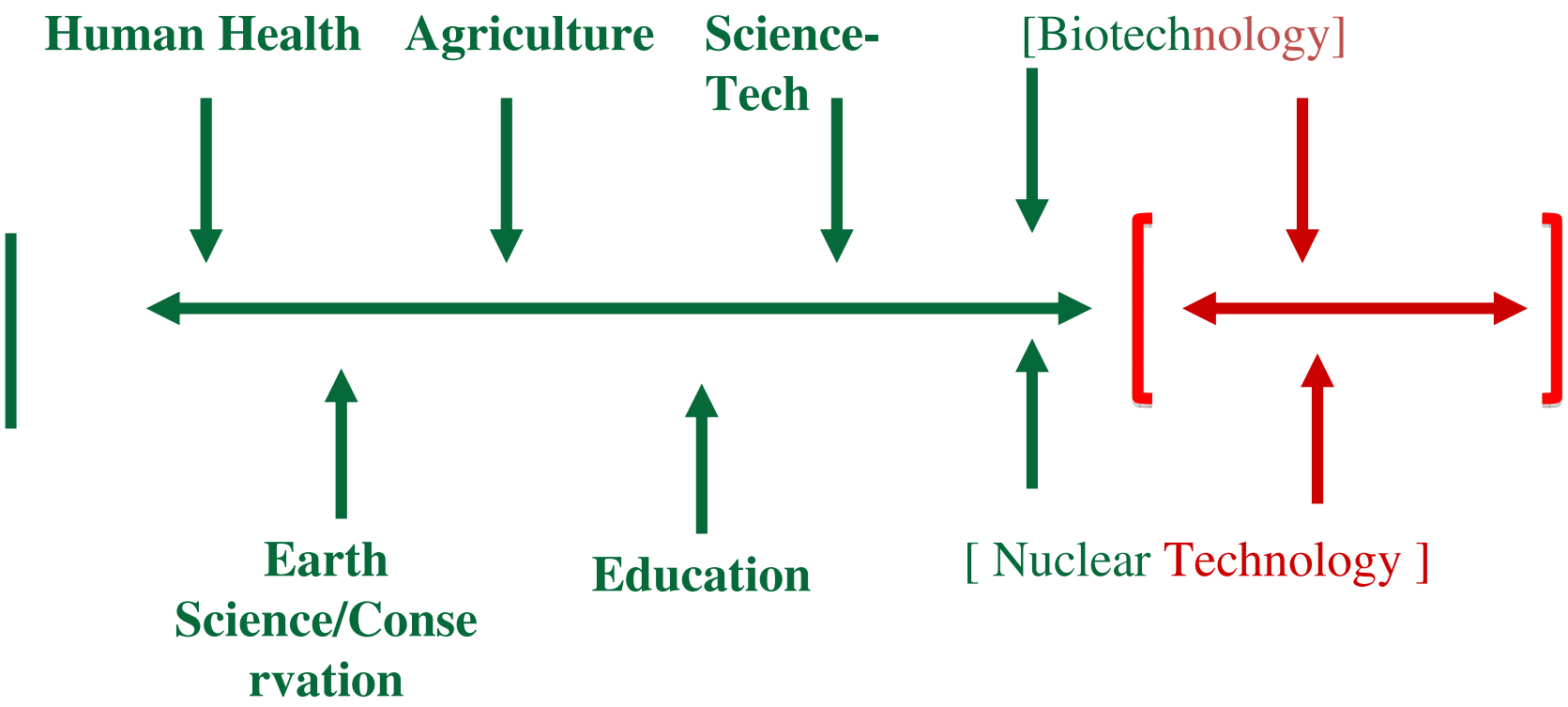
Polemically / politically there is a spectrum of public welfare that argues that much – perhaps not all ? -- data should be released?

OR perhaps *all of it* should be released???

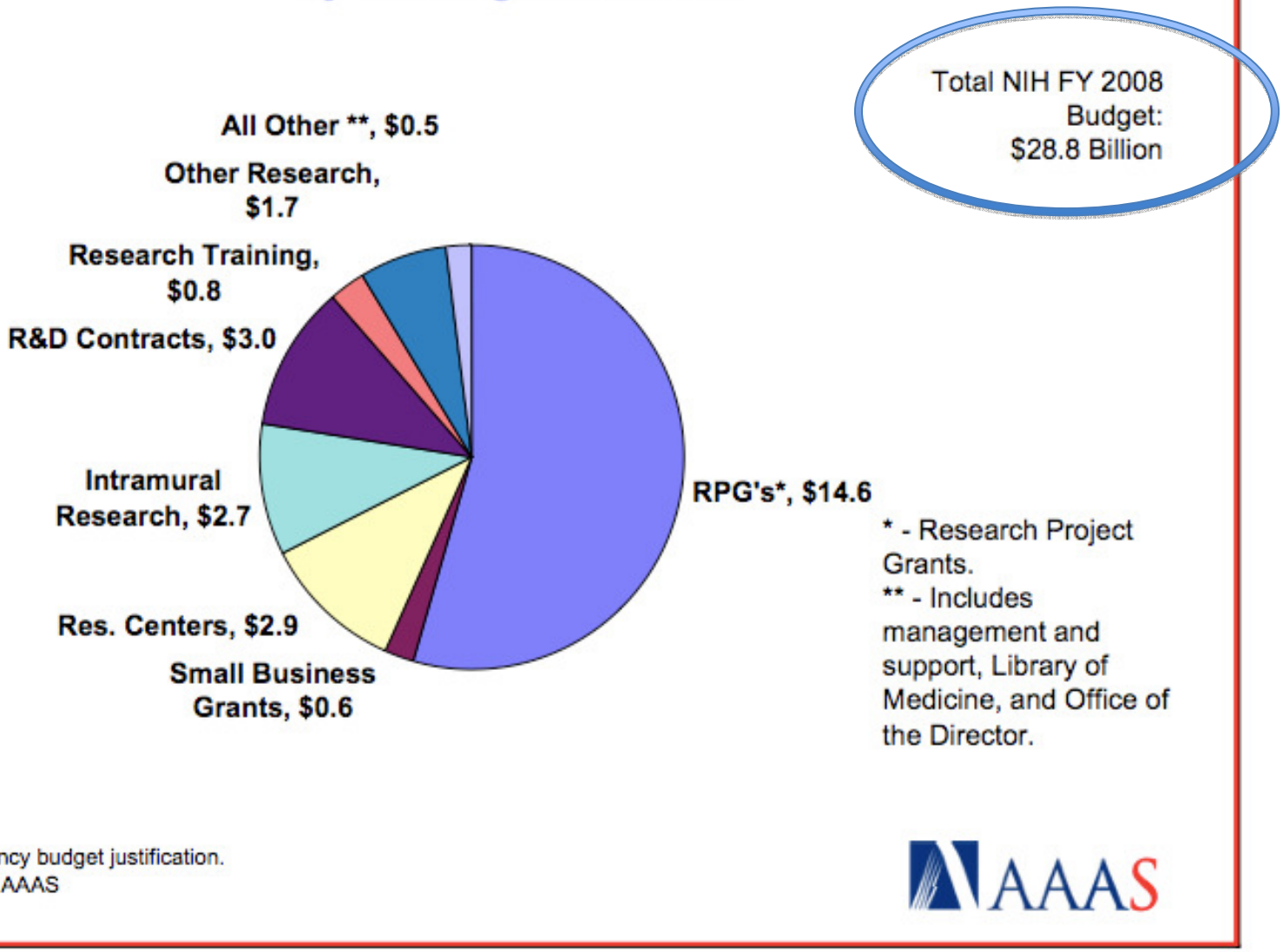
- Consider the Faustian / Klaus Fuchs / Abdul Qadeer Khan syndrome – see NY Review of Books: Volume 56, Number 6 · April 9, 2009 Jeremy Bernstein, *He Changed History)]*

Note how many of Thursday's arguments focused on human health and welfare– it is the easiest / most obvious case

<u>ALL</u> knowledge? Or perhaps, an ethical spectrum ? – the polemics of support for the Science Knowledge Commons



National Institutes of Health FY 2008 Budget by Funding Mechanism


Total NIH FY 2008 Budget: $28.8 Billion
All Other **, $0.5
Other Research, $1.7
Research Training, $0.8
R&D Contracts, $3.0
Intramural Research, $2.7
Res. Centers, $2.9
Small Business Grants, $0.6
RPG's*, $14.6
* - Research Project Grants.
** - Includes management and support, Library of Medicine, and Office of the Director.

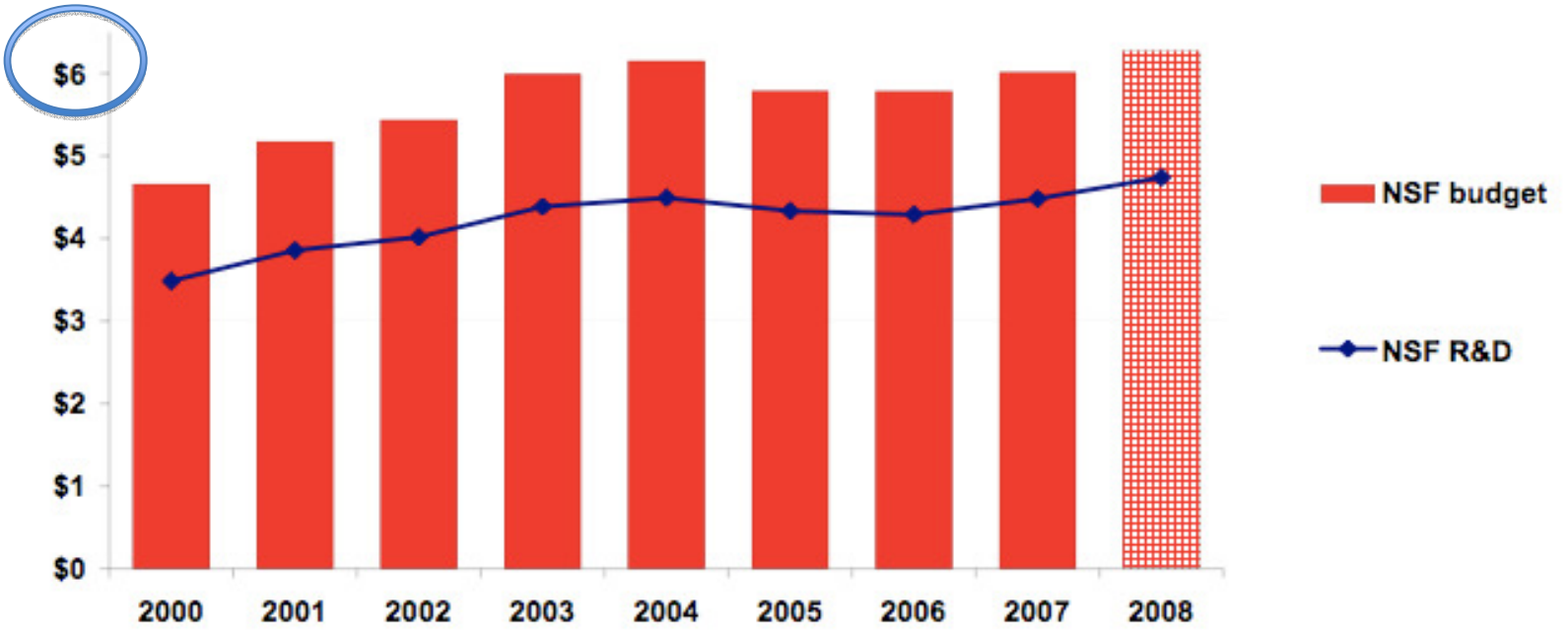
Source: NIH agency budget justification.
FEB. '07 © 2007 AAAS

AAAS

http://www.aaas.org/spp/rd/fy08.htm

National Science Foundation Budget, FY 2000-2008

(budget authority in billions of constant FY 2007 dollars)



Source: National Science Foundation, and latest AAAS estimates of FY 2008 budget. FY 2008 is budget request; FY 2007 is estimate of final appropriation.

FEB. '07 REVISED © 2007 AAAS

AAAS

http://www.aaas.org/spp/rd/fy08.htm

- The global community is focusing on full-life-cycle management of data
  - Particularly including curation and preservation [bit rot ?]
    - Migration? / Emulation?
    - Trusted Digital repositories

Migration / Emulation ???

8”


3M
Wang
compatible
SS,SD,32RH
single side
single density
32 sectors
Up
Insert
WRITE PROTECT
This Diskette has the capability of being write protected by the hole indicated by the arrow below. When the hole is open, it is protected; when covered, writing is allowed. The hole is closed by placing a tab over the front of the hole, and folded over covering the rear of the hole. The Diskette can then be write protected by removing the tab.

Preservation



objet trouvé – gutter, 10th & Colorado, Santa Monica



Internet Archive

- A focus on broadband alone is not sufficient
- The technology for affordable mass digitization exists – should be part of any economic stimulus effort
  - Library of Congress scanning center / FedLink eligibility for US Federal governmental contracts
  - WayBackMachine / Archive-It/ Internet Archive
    - 150 billion Web pages
  - NASA Images Project

Wednesday, January 21st, 2009 at 12:00 am
Freedom of Information Act

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

SUBJECT: Freedom of Information Act

A democracy requires accountability, and accountability requires transparency. As Justice Louis Brandeis wrote, "sunlight is said to be the best of disinfectants." In our democracy, the Freedom of Information Act (FOIA), which encourages accountability through transparency, is the most prominent expression of a profound national commitment to ensuring an open Government. At the heart of that commitment is the idea that accountability is in the interest of the Government and the citizenry alike.

The Freedom of Information Act should be administered with ***a clear presumption: In the face of doubt, openness prevails.*** The Government should not keep information confidential merely because public officials might be embarrassed by disclosure, because errors and failures might be revealed, or because of speculative or abstract fears. Nondisclosure should never be based on an effort to protect the personal interests of Government officials at the expense of those they are supposed to serve. In responding to requests under the FOIA, executive branch agencies (agencies) should act promptly and in a spirit of cooperation, recognizing that such agencies are servants of the public.

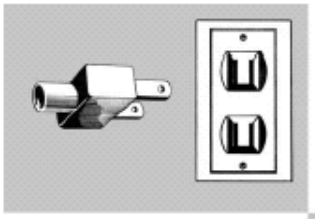
All agencies should adopt a presumption in favor of disclosure, in order to renew their commitment to the principles embodied in FOIA, and to usher in a new era of open Government. ***The presumption of disclosure should be applied to all decisions involving FOIA.***

http://www.whitehouse.gov/the_press_office/Freedom_of_Information_Act/

Electric Current Abroad

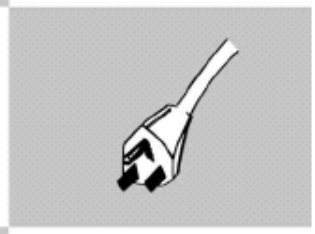
U.S. Department of Commerce
International Trade Administration
February 2002

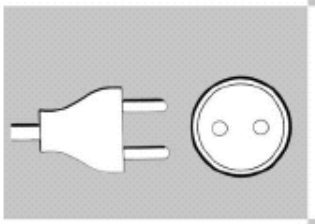
Plugs in Commercial Use



Type
A
Flat blade attachment plug

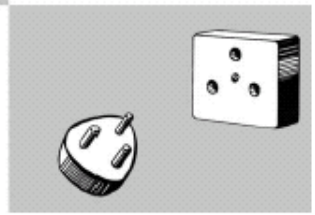
Type
B
Flat blades with round grounding pin

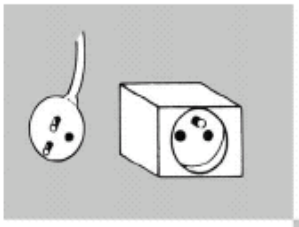




Type
C
Round pin attachment plug

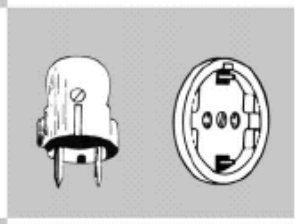
Type
D
Round pins with ground

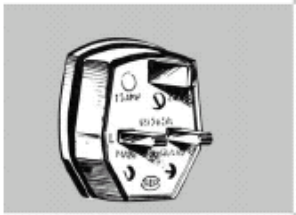




Type
E
Round pin plug and receptacle with male grounding pin

Type
F
"Schuko" plug and receptacle with side grounding contacts

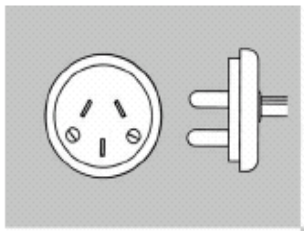




Type
G
Rectangular blade plug

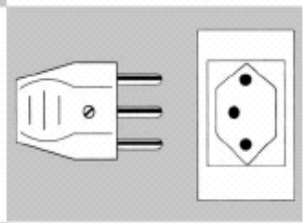
Type
H
Oblique flat blades with ground

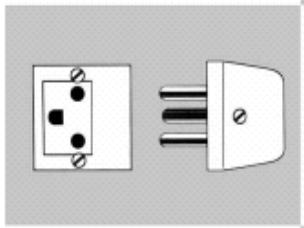




Type
I
Oblique flat blade with ground

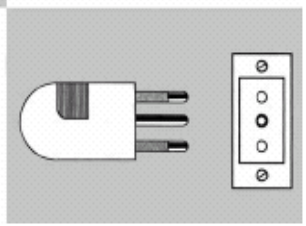
Type
J
Round pins with ground





Type
K
Round pins with ground

Type
L
Round pins with ground





http://www.mikero.com/blog/2009/02/20/more-darwin
http://www.zazzle.com/darwin2009

Tom Moritz

Internet Archive

+1 310 963 0199

<moritz@archive.org>

